Water Resources Research

RESEARCH ARTICLE
10.1029/2018WR023989

Key Points:
- Diagnosing model deficiency is difficult as there are many possible causes of poor simulations
- Split sample failure does not mean model structural inadequacy—further tests are required
- This framework diagnoses the cause of poor performance and prioritizes remedial action

Supporting Information:
- Supporting Information S1

Correspondence to:
K. Fowler,
fowler.k@unimelb.edu.au

Citation:
Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., et al. (2018). Simulating runoff under changing climatic conditions: A framework for model improvement. *Water Resources Research*, *54*. https://doi.org/10.1029/2018WR023989

Received 28 AUG 2018
Accepted 16 SEP 2018
Accepted article online 1 OCT 2018

Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement

Keirnan Fowler[1], Gemma Coxon[2,3], Jim Freer[2,3], Murray Peel[1], Thorsten Wagener[3,4], Andrew Western[1], Ross Woods[3,4], and Lu Zhang[5]

[1]Department of Infrastructure Engineering, University of Melbourne, Parkville, Victoria, Australia, [2]School of Geographical Sciences, University of Bristol, Bristol, UK, [3]Cabot Institute, University of Bristol, Bristol, UK, [4]Department of Civil Engineering, University of Bristol, Bristol, UK, [5]CSIRO Land and Water, Canberra, ACT, Australia

Abstract Rainfall-runoff models are often deficient under changing climatic conditions, yet almost no recent studies propose new or improved model structures, instead focusing on model intercomparison, input sensitivity, and/or quantification of uncertainty. This paucity of progress in model development is (in part) due to the difficulty of distinguishing which cases of model failure are truly caused by structural inadequacy. Here we propose a new framework to diagnose the salient cause of poor model performance in changing climate conditions, be it structural inadequacy, poor parameterization, or data errors. The framework can be applied to a single catchment, although larger samples of catchments are helpful to generalize and/or cross-check results. To generate a diagnosis, multiple historic periods with contrasting climate are defined, and the limits of model robustness and flexibility are explored over each period separately and for all periods together. Numerous data-based checks also supplement the results. Using a case study catchment from Australia, improved inference of structural failure and clearer evaluation of model structural improvements are demonstrated. This framework enables future studies to (i) identify cases where poor simulations are due to poor calibration methods or data errors, remediating these cases without recourse to structural changes; and (ii) use the remaining cases to gain greater clarity into what structural changes are needed to improve model performance in changing climate.

Plain Language Summary Rainfall runoff models are tools used by hydrologists in climate change assessments to estimate how future streamflow might change in response to a given (often hypothetical) climate scenario. For example, suppose we can assume that rainfall in a particular location is going to reduce by 20% in the future. Does this mean that streamflow will also reduce by 20%? Or will it be 10% less or 40% less? Although rainfall runoff models are among the best tools available, they are often not very good at answering this question. When tested on historical multiyear droughts, they often perform poorly, and we are unsure why. One problem is that when a model fails in this task, it is difficult to know what went wrong. Perhaps there was a problem with the data, since environmental monitoring is often subject to large errors. Perhaps the problem lay not with the model itself but with the way it was trained, or calibrated, to the data. Lastly, perhaps the model itself—its mathematical equations—need to be changed. To improve our estimates, we need a method to test which cause is behind the model failure; otherwise, we might make changes where none are warranted. This paper proposes such a method, in the form of a multistep framework that can isolate the causes of model failures. By ensuring that our attention is focused in the correct direction, this framework will help us to understand and make better estimates of how river flow will be altered by a changing climate.

1. Introduction

For effective water resource planning under climate change, it is essential to understand how catchments respond to changes in climatic forcing. Future climatic changes may go beyond the variability of the past (Covey et al., 2003; Forster et al., 2007; Meehl et al., 2007; Milly et al., 2008) and be amplified by hydrologic systems (e.g., Saft et al., 2015; van Dijk et al., 2013; Vogel et al., 1999), so it is important to learn what we can from past climate sequences and ensure hydrological models are improved by this learning. Such models are key tools for quantifying rainfall-runoff responses to climate model projections (e.g., Bergström et al., 2001; Chiew et al., 1995, 2009; Christensen et al., 2004; Faramarzi et al., 2013; Forzieri et al., 2014; Krysanova et al., 2017; Pechlivanidis et al., 2017; Samaniego et al., 2017; Singh et al., 2014; Smith, Bates

et al., 2014; Smith, Freer et al., 2014), informing planning of possible responses and/or adaptation. It is therefore critical to ensure rainfall-runoff models can provide robust simulations under changing climatic forcing, in line with the wider current International Association of Hydrological Sciences emphasis on *change in hydrology and society* (Montanari et al., 2013).

In practice, rainfall-runoff models often show significant reductions in performance when applied in climatic conditions different to the calibration data. Vaze et al. (2010) tested four rainfall-runoff model structures on 61 catchments in Australia and reported that model performance tended to decline in proportion to the change in climatic variables between calibration and validation period, which has been confirmed by many other studies (e.g., Broderick et al., 2016; Coron et al., 2012, 2014; Freer et al., 2003; Refsgaard & Knudsen, 1996; Saft et al., 2016; Seiller & Anctil, 2015; Seiller et al., 2012). A related problem is that model parameters, classically thought of as representing time-invariant properties of river catchments, usually vary depending on which time period a model is calibrated to. Merz et al. (2011) reported that these changes are strongly related to the average climatic conditions (e.g., temperature) in the calibration period (see also Brigode et al., 2013; de Vos et al., 2010; Freer et al., 2003; Wilby, 2005). The nonstationarity of calibrated parameter values may be a form of compensation for models that are missing key processes (or constraining them poorly) occurring in catchments that are subjected to long-term drying or wetting, with poor validation performance the ultimate outcome (Beck, 2002; Wagener, 2003).

Faced with rainfall-runoff models that are often deficient under changing climatic conditions, there is a clear need for new or improved models (and modeling methods) that are more robust. However, very little progress has been made toward improving models, or creating new ones, for better robustness under changing climatic conditions. To illustrate this, a literature review was conducted to identify every study with a DOI that cited both of the studies mentioned in the previous paragraph—namely, Vaze et al. (2010) and Merz et al. (2011). Of the 55 studies fulfilling these criteria, only one (Westra et al., 2014) resulted in a new model structure. The most common topics were climate change impact assessments and associated methods (10 studies), papers examining input or output uncertainty or sensitivity (7 studies), model intercomparison (6 studies), and papers proposing improvements to modeling practice (e.g., improved calibration or validation procedures; 6 studies). Each of these topics are individually important and worthy contributions to the literature. However, the collective lack of studies proposing new or improved models is both striking and concerning and is particularly surprising given the proliferation of software schemes that facilitate comparison of alternative model structures, such as FUSE (Clark et al., 2008), SUPERFLEX (Fenicia et al., 2011) and SUMMA (Clark et al., 2015), and others that can choose between candidate structures (e.g., Marshall et al., 2007). Note that under the phrase *new or improved models* we include systematic approaches to temporal changes in parameters and explicit inclusion of these in modeling frameworks (e.g., Kelleher & Shaw, 2018).

One reason for the paucity of progress in model development is that structural inadequacy is only one cause of poor simulations, with confounding factors such as deficient calibration methods and/or data errors also contributing to simulation errors (Beven et al., 2011; Beven & Westerberg, 2011; Coron et al., 2014; Kavetski et al., 2006; McMillan et al., 2012; Seibert, 2003; Singh et al., 2011). Thus, the problem is initially one of diagnosing the salient cause of model failure, and this paper presents a framework to do this. While recognizing that any given case of poor performance may be due to a mixture of reasons, the tests described herein provide a rational basis to prioritize remedial action, which could include (i) improving the calibration method; (ii) rectifying data errors; and (iii) model structural improvements (i.e., changes to model equations). Our view is that changes to model equations are appropriate only after conducting basic tests to detect data errors, and assessing the possibility of poor parameterization, and we suggest methods herein for each purpose.

This framework is significant because it complements existing literature concerning rainfall-runoff model assessment and improvement. We argue that the following three types of framework are required to improve rainfall-runoff simulations in changing climate: (1) frameworks to test model validation performance (examples include Coron et al., 2012; Klemeš, 1986; Seibert, 2000 and Thirel et al., 2015b); (2) frameworks to determine the salient cause of poor validation performance—this is a gap in the current literature, which this paper seeks to address; and (3) if the cause is confirmed to be deficient model structure, frameworks to detect and fix model structural issues (examples include Gupta et al., 2008; Westra et al., 2014). Therefore, the framework presented herein sits between existing schemes, and together with them it completes the workflow that modelers require to move from split sample testing through to model structural improvement.

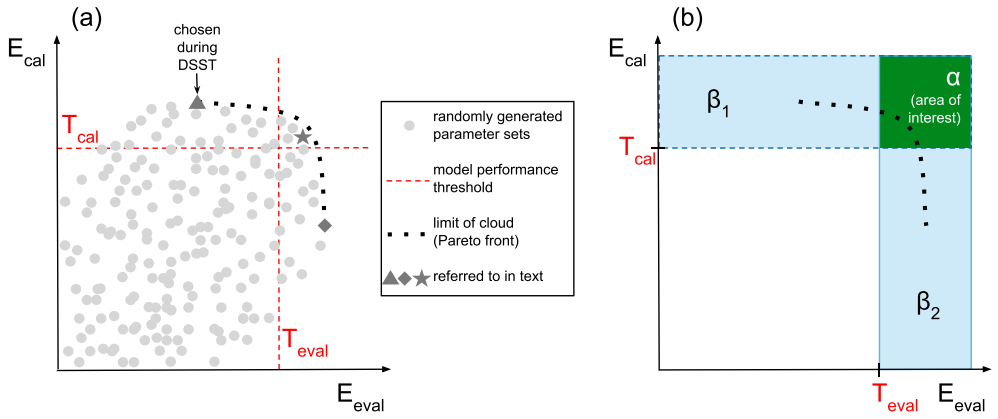


Figure 1. (a) Conceptual diagram of model performance E for a randomly generated ensemble of parameter sets, along with modeling performance acceptance thresholds T for calibration and evaluation periods (with contrasting climate). (b) Demarcation of the space into regions β_1, β_2, and α, where $\alpha = \beta_1 \cap \beta_2$. Note that the scheme is generalizable for cases where more than one evaluation period is used (i.e., for N evaluation periods, $\alpha = \beta_1 \cap \beta_2 \cap \ldots \cap \beta_{N+1}$). DSST = Differential Split Sample Test.

2. Rationale of Framework

In this section the logical basis of the framework is explained. First, we define notation: Let model performance (by whatever measure) be denoted by E, and let T be some threshold of performance that defines model adequacy—that is, T is a minimum acceptable value of E. E and T are defined separately for the calibration period and for an evaluation period with contrasting climate (Klemeš, 1986, Test 2a). Note we hereafter avoid the term *validation* because models of open systems can never be validated (Oreskes et al., 1994); we use the term *evaluation* instead. We recognize that there is more to split sample testing than a simple comparison of performance metrics (see section 5.3 and Gupta et al., 1998; Yapo et al., 1996) but for the present section we seek to present the rationale for the framework in the simplest possible terms.

In the literature, the most common outcome is that models fail split sample testing during evaluation. This means that the parameter set chosen during the split sample test meets the criteria $E_{cal} > T_{cal}$ but fails the other criteria (i.e., $E_{eval} < T_{eval}$). Expressed diagrammatically, this parameter set is the triangle in Figure 1a. However, the parameter set(s) chosen during split sample testing does not necessarily represent the full capabilities of the model structure (Fowler et al., 2016). This can be revealed by considering an ensemble of randomly generated parameter sets, shown in Figure 1a as gray dots. For example, parameter sets are available with $E_{eval} > T_{eval}$ (e.g., the diamond). Also, other parameter sets fulfill both $E_{cal} > T_{cal}$ and $E_{eval} > T_{eval}$ simultaneously (e.g., the star), which indicates that no changes to the model structure are required to meet both acceptance thresholds, but the wrong parameter set was initially chosen during the split sample test. In this case, the mathematical optimum during calibration (triangle, highest E_{cal} value) is not a hydrological optimum (e.g., star) where more consistent performance in both time periods is found (e.g., Andréassian et al., 2012).

Figure 1a is a conceptual example only—in practice, the coverage of parameter sets in the space will vary for each case, and a key task of the framework is to determine this coverage. The framework uses the geometry of the coverage to prioritize remedial action in cases of split sample failure. T_{cal} and T_{eval} are used as dividing lines through the 2-D space (Figure 1b) to create regions β_1 and β_2, respectively. An additional region α is defined as $\beta_1 \cap \beta_2$. Coverage over these areas provides valuable information to guide remedial action, such as the following:

- In the case of coverage of α (Figure 2a), no changes to the model structure are required to meet both acceptance thresholds. The priority should be to improve the calibration method, so that when the Differential Split Sample Test (DSST) is repeated a more suitable parameter set (or sets) is identified;
- For coverage of β_1 and β_2 but not α (Figure 2b), the model structure is flexible (it can fit data it is calibrated to) but not transferable (good parameter sets in one period are poor in others), a situation best remedied by model structural changes;

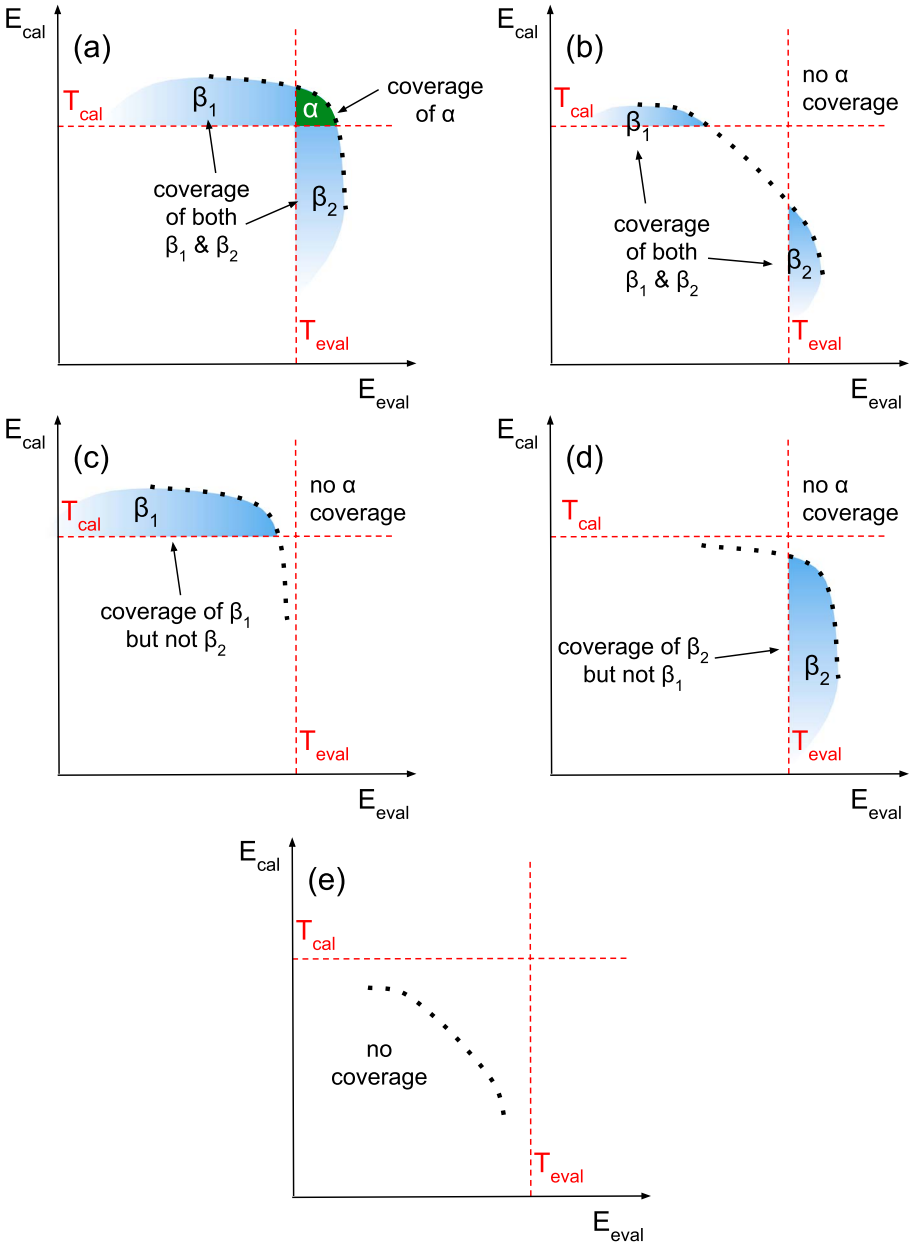


Figure 2. Results categories when testing the coverage created by a random ensemble of parameter sets, in two-dimensional E_{cal}-E_{eval} space. Panel (a) depicts the case where the α and both β regions have coverage, as distinct from (b) coverage of both β regions but not the α region, (c, d) one β coverage but not the other, and (e) no coverage of any region. Categories are used to inform courses of action in investigating and remediating causes of poor performance.

- If β_1 coverage exists but β_2 does not (or vice versa; Figures 2c and 2d), something is impeding performance in one of the periods more than the other, which may indicate a temporal trend in data errors or a model structure more suited to one set of climatic conditions than another. Further tests are required to distinguish between these causes (see below); and
- In the case of coverage of neither β_1 nor β_2 (Figure 2e), something is impeding performance over both climatic periods, and there is a relatively higher chance that data errors are present (which can be tested separately as described below) or the model may be inadequate in a way that varies little with climatic conditions.

The core tasks of the framework are to (a) categorize the coverage as per the above and (b) undertake further tests aimed at confirming and supplementing the diagnosis suggested by the coverage, namely,

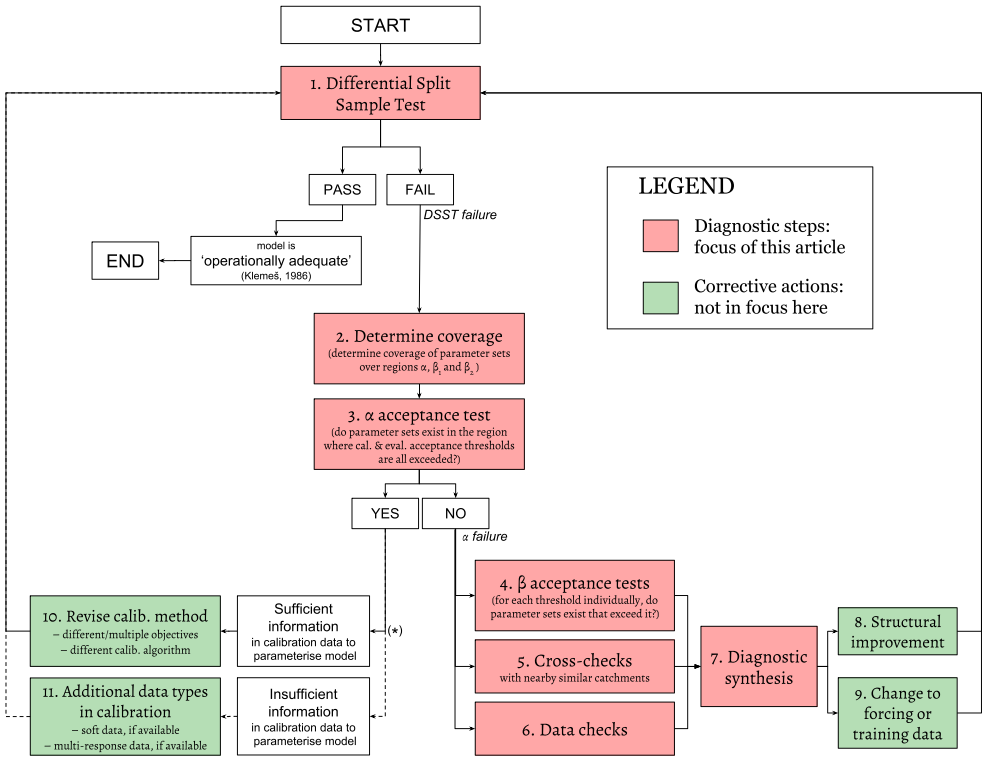


Figure 3. Flowchart showing steps to apply framework. The modeler proceeds through the steps based on the results of diagnostic tests (colored red), except for the decision point at (*) which is difficult to know a priori (see step 10). DSST = Differential Split Sample Test.

- *Multicatchment cross checks* that repeat the above test in similar nearby catchments. A comparable coverage strengthens the initial diagnosis, whereas dissimilar coverage may reveal one-off error sources such as observation errors that affect one catchment but not others; and
- *Data-focused tests* to identify data errors, for example, based on temporal consistency of different data.

Together, these tests have sufficient diagnostic power to prioritize remedial action in most cases of DSST failure. For each test, further detail is provided in subsequent sections.

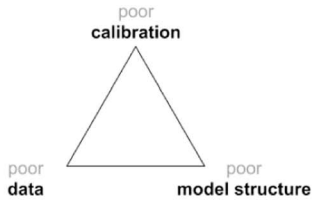
3. Step-by-Step Description of Framework

Figure 3 gives a flow chart showing the steps needed to apply the framework. This paper focuses on the diagnostic steps (shown in red), rather than the corrective actions (shown in green). Many corrective actions can be undertaken using existing techniques, as noted in the text where relevant. The framework is most applicable to cases of DSST failure—to make this explicit, we include the DSST as the first step.

3.1. Diagnostic Steps

1. *DSST*: This is Test 2a from Klemeš (1986), involving these steps: (a) *divide* the historic record into two or more periods with contrasting climate; (b) *calibrate* the rainfall-runoff model on one of the periods and evaluate on the other(s); (c) *quantify* model performance with some numerical measure *E*; and (d) *evaluate* performance for each period by checking whether *E* exceeds acceptance thresholds *T*, for each period. As discussed in the supporting information, Text S1:

- *E*, the acceptance metric, should reflect the modeling purpose. Multiple acceptance metrics may be used if appropriate (e.g., Thirel et al., 2015);
- *T*, the thresholds of acceptance, should be defined a priori and preferably reflect the modeler's understanding of the level of accuracy required for decision making;
- The calibration objective function need not correspond to *E*. Robust simulation performance in changing climatic conditions depends on fidelity of process representation, so calibration objective function(s)



Example	Meaning: supports diagnosis of...
	poor calibration
	poor data
	poor data or model structure, but not poor calibration

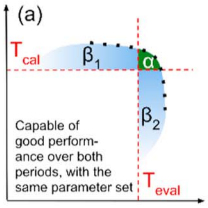
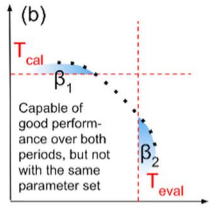
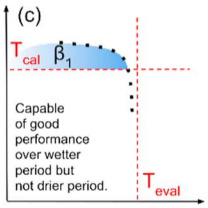
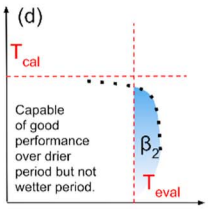
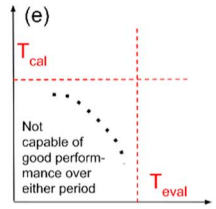
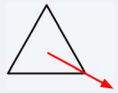
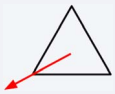
Coverage	Diagnosis notes	
(a) 	 Parameter sets that meet acceptance thresholds do exist within the model structure. In this case, the priority is improving the calibration method, so that when the DSST is repeated a more suitable parameter set is identified.	
(b) 	 The model structure is flexible (it can fit data it is calibrated to) but not transferable (good parameter sets in one period are poor in others). Focus on model structure: given the absence of performance problems in each period individually, it is relatively less likely (but not impossible) that data errors are the salient cause.	
	Is the same result seen in nearby similar catchments[1]?	
	No	Yes
(c)  (d) 	 Something is impeding good performance in one of the periods more than the other. Given it is not shared with nearby similar catchments, it is relatively more likely to be a data issue specific to this catchment.	 Something is impeding good performance in one of the periods more than the other, consistently across multiple catchments. Further testing (Table 1) may reveal whether it is a data or model structural issue.
(e) 	 A structural deficiency Would likely cause consistently poor Pareto results across multiple similar catchments. Thus, this result suggests a data issue specific to this catchment.	 Except in cases of systematically poor data across multiple catchments[2], repeated type (e) failure indicates a generally poor[3] model structure.

Figure 4. Possible outcomes in the α and β coverage tests, for the case of one acceptance metric for calibration and one for evaluation. Note that this figure assumes the evaluation period is drier than the calibration period, as per the example in section 4. Footnotes: [1]Assuming the existence of such, since nearby catchments may be hydrologically different. [2]That is, cases where an entire data collection regime is systematically flawed. [3] *Generally poor* in this case means poor regardless of climatic period.

Table 1
Data Checks Undertaken at Step 6

Error category	Examples of specific cause	Test ID	Data check/test	Diagnosis If yes	Diagnosis If no
Temporal changes/trends in precipitation error	Decommissioning of rain gauge affecting interpolations; or tree growth obstructing rain near gauge	i	Does a double mass curve (with nearby independent precipitation data) have a change in gradient?		
Event-based precipitation errors	Runoff generating storms missed by local rainfall gauges	ii	Can missed events (or added events) be detected via time series/single mass curve comparisons between precipitation and runoff?		
Consistent long-term bias in precipitation	Rain gauge not representative of catchment-average precipitation; or deficiency in spatial interpolation	iii	Part 1: Review available information on the precipitation or runoff data. If errors seem likely, conduct the following test:		If no to all data checks that are conducted, 
Rating curve error, temporally consistent	Stream gauge is poorly rated, causing permanent bias		Part 2: Does adding a calibratable, temporally constant scaling factor to precipitation allow coverage of the α region? Or at least, significantly improve performance?		
Rating curve error, temporally changing	Erosion around stream gauge	iv	Does a double mass curve (with nearby independent runoff data) have a change in gradient?		
Inappropriate PET formulation	Using a formulation that ignores wind in an area subject to trends in wind	v	Part 1: Is there a long-term trend in a variable that is not included in the PET formulation? If yes, revise PET formulation accordingly, and Part 2: Does trialing the revised PET formulation allow coverage of the α region? Or at least, significantly improve performance?		

Note. PET = potential evapotranspiration.

should be chosen to extract information relevant to these processes from the calibration data (Fowler et al., 2018; Gupta et al., 1998);
- More than one evaluation period may be used (Coron et al., 2012; Thirel et al., 2015);
- The calibration algorithm could be optimization or an ensemble method (e.g., Beven & Binley, 1992; Vrugt et al., 2008). The latter may be used so long as the experimental design yields an objective DSST result.

All subsequent steps of the framework are for cases where the model fails the DSST.

2. *Determine coverage over regions α, β_1, and β_2*: Coverage can be determined by either generating a random ensemble of parameter sets, as per Figure 1a, or using a multiobjective optimizer to generate the Pareto front (dotted line in Figure 1) between E_{cal} and E_{eval}. While a random ensemble is simpler, an impractically large number of parameter sets may be needed to achieve sufficient sampling density (Fowler, 2017, Figure C.4). Thus, a multiobjective optimizer is recommended (e.g., Hadka & Reed, 2013; Vrugt & Robinson, 2007). The full range of possible coverage results, along with associated diagnoses, are given in Figure 4.
3. *The α acceptance test*: The model structure passes this test if the model structure has coverage of the α region (i.e., if parameter sets exist there; Figure 4a) and fails otherwise. A pass means that no changes to the model structure are required to exceed all acceptance thresholds, even though the parameter set chosen at step 1 did not do so. For an α pass, the suggested remedial action is to improve the calibration method used in the DSST so that a more suitable parameter set (or sets) can be identified in a repeat DSST. Model structures failing the α test cannot meet the thresholds no matter how they are calibrated, so it is pointless to focus on DSST calibration method, and the only options are to consider possible data errors or structural deficiency.

4. *The β acceptance tests*: One β test is undertaken for each acceptance metric. The result is a pass if the model structure has coverage of the corresponding β region, and a fail otherwise. β results are interpreted in step 7 (cf. Figure 4).
5. *Cross checks with other catchments*: In this step, the α and β acceptance tests are repeated in nearby similar catchments. Although highly recommended, this step depends on data availability and is not always possible, particularly if nearby catchments are not hydrologically similar or are ungauged. The confirmation (or not) of acceptance test results in nearby similar catchments allows considerable diagnostic insight (see Figure 4 and step 7), and analysis across larger samples may lead to improved process understanding in its own right (Gupta et al., 2014).
6. *Data checks*: Table 1 lists checks that can be undertaken at this step. For tests with two parts (iii and v), only the first part may be required. Although some tests involve simulation, these tests are still categorized as *data checks* because of their focus on data quality/suitability.
7. *Diagnostic synthesis*: In this step, the various test outcomes are interpreted and used jointly to deliver a diagnosis; that is, the β coverage test outcomes are used together with multicatchment cross checks and data checks to develop greater weight of evidence.

3.2. Corrective Actions

The corrective actions are listed here using the step numbers from Figure 3. Note that, unlike earlier steps, steps 8–11 are not sequential. Rather, they are undertaken based on results of steps 1–7, as per the arrows in Figure 3 and the triangular diagrams in Figure 4 and Table 1.

8. *Structural improvement* (triangular diagram with arrow facing down and to the right in Figure 4 and Table 1) means changing one or more of a model's equations. Methods to formulate structural improvement are not covered here because, as noted, the present focus is on diagnostic tests rather than remedial actions. However, structural improvements are discussed by, for example, Ambroise et al. (1996), Beck (2005), Gupta et al. (2008), de Vos et al. (2010), Westra et al. (2014), and Kelleher and Shaw (2018).
9. *Changes to forcing or training data* (triangular diagram with arrow facing down and to the left in Figure 4 and Table 1): An important clarification is that this step does not involve changing the *type* of forcing or training data. Rather, this step involves fixing, if possible, a problem identified with the existing data. Examples include reinterpolating precipitation inputs after removing a problematic rain gauge, excluding from calibration/evaluation those parts of the record known to contain significant errors in flow gauging, or adopting a more appropriate formulation of potential evapotranspiration (PET).
10. *Revising calibration method* (triangular diagram with arrow facing up in Figure 4 and Table 1): Having confirmed coverage of the α region, the focus should be on finding a parameter set in this region using data from the calibration period only, via an improved DSST calibration method. Potential options here include changes in objective function (Fowler et al., 2018), a different calibration algorithm, multiobjective methods including subperiod analysis (e.g., Choi & Beven, 2007; Freer et al., 2003; Gharari et al., 2013), and limits of acceptability methods (Beven, 2006). *Trial and error* should be avoided by choosing methods rationally with recourse to hydrological theory (i.e., *why* would we expect an improved outcome?).

When seeking to improve calibration methods, modelers should be mindful that there may not be sufficient information in the calibration data to parameterize the model. If processes that are inactive or unimportant during the calibration period become important when climatic conditions change, it may be unreasonable to assume that the model parameters that govern such processes are identifiable based on prechange data alone (Andréassian et al., 2012; Ljung, 1998; Reichert & Omlin, 1997; see van Werkhoven et al., 2008 for a demonstration of this issue across space). Alternatively, information regarding these processes may be present in the calibration period in nondischarge data types (e.g., groundwater, soil moisture, and remotely sensed data), so calibration strategies that use these data may aid identifiability of key parameters. This is difficult to know a priori, so it is suggested to initially seek an improved calibration method with the data at hand (step 10), introducing additional data types later as required and if available (step 11).

11. *Additional data types in calibration*: Additional data types could include observations of groundwater, water quality, soil moisture, vegetation data, actual evapotranspiration (AET), snow data, qualitative or fuzzy measures that reflect subjective understanding of dominant processes, or any other data type

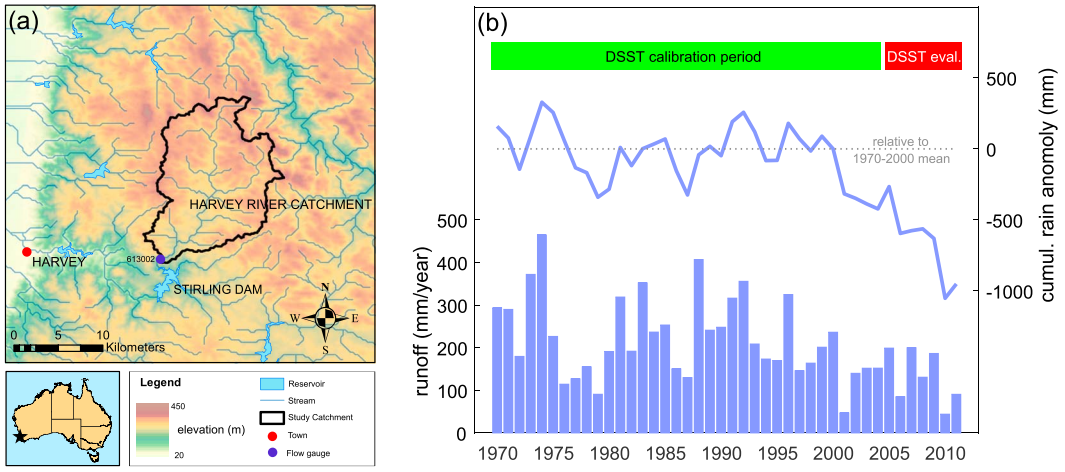


Figure 5. (a) map of the study catchment, Harvey River at Dingo Road in southwest Australia (613002). (b) Annual streamflow (columns) and cumulative rainfall anomaly relative to the 1970–2000 mean (line), in addition to periods used for model calibration and evaluation in the Differential Split Sample Test (DSST).

that may aid model identification (see supporting information Text S2). This step may be also undertaken to increase the realism of the model (Clark et al., 2011; Freer et al., 2004; Seibert & McDonnell, 2002).

4. Example Application

4.1. Example Catchment and Data

This section describes a single-catchment case study, taking the perspective of a climate change impact assessment. Although the framework is also applicable to large catchment samples (supporting information Text S10), for clarity of presentation we focus on one catchment only. The aim is a calibrated rainfall-runoff model suitable for water resources assessment under projected future climatic conditions. The selected catchment is Harvey River in the south west of Australia (Figure 5a), upstream of the gauge at Dingo Road (station 613002, area 148 km^2, mean annual rainfall 1,000 mm/year, runoff ratio 0.2) which is one of Australia's *Hydrologic Reference Stations* (Turner, 2012). Downstream of the study area, the river flows into the 57 GL Stirling Reservoir, used for metropolitan supply. The catchment has a strongly seasonal climate with hot dry summers and cool, wet (but snow-free) winters, with >80% of flow occurring in the months of July to November. High sensitivity of runoff to slight differences in long-term rainfall is typical of this area due to nonlinear groundwater-surface water interactions (Hughes et al., 2012; Kinal & Stoneman, 2012). Interannual variability in rainfall is high, and a run of low-rainfall years in the 2000s led to a cumulative rainfall deficit (Figure 5b), relative to earlier decades. The associated region-wide decline in streamflow led to construction of a desalination plant for Perth and a greater dependence on groundwater (Petrone et al., 2010). Human impacts are minimal since the study catchment is entirely forested (mostly Jarrah hardwoods, *Eucalyptus marginata*), with no temporal changes to land use. No major bushfires have affected the catchment over the hydrometeorological record.

A lumped modeling approach with a daily time step is adopted, with daily catchment average rainfall derived from the interpolated (~5 km) gridded rainfall product of Jones et al. (2009), and PET estimates derived according to the Wet Environment method from Morton (1983), using the gridded data set produced by Jeffrey et al. (2001) extracted at the catchment centroid. The extracted PET has an average annual value of 1,340 mm/year, approximately 35% greater than rainfall. Catchment boundaries are derived using D8 flow analysis on postprocessed Shuttle Radar Topography Mission data published by Gallant et al. (2011), on a grid size of 1 s (approximately 30 m).

4.2. Example Model Structures

Since the focus is on the framework itself, this paper presents no new model structures. Instead, we choose a model structure which has two preexisting variants that provide a case study in model improvement, namely, IHACRES. Here both versions have a daily time step and are spatially lumped.

The first version, termed IHACRES-A in this paper, follows the descriptions of Jakeman et al. (1990) and Jakeman and Hornberger (1993) with six free parameters. A schematic is provided in the supporting information Figure S1. IHACRES-A implements an index of catchment wetness that increases linearly in response to rainfall (as a function of parameter c) and decreases nonlinearly in response to PET (as a function of parameters c, T_w, and f). The wetness value determines the proportion of rainfall converted to runoff. Runoff is routed through two parallel linear stores with different time constants (parameters T_q and T_s). The split between the routing stores is determined by parameter V_s. Model parameters and thresholds are provided in supporting information Table S1.

The second version, IHACRES-B, includes changes by Ye et al. (1997), who retained the index of catchment wetness but allowed for a nonlinear relationship (described by new parameter p, $p \geq 1$) between the index and runoff proportion. They also enforced a threshold on index values (set by new parameter l, $l \geq 0$) which must be exceeded before any rainfall can be converted to runoff. Since they worked in ephemeral catchments, Ye et al. (1997) removed the slower of the two routing storages. However, given Harvey River flow is relatively sustained and historically perennial, it is appropriate to retain both routing storages in IHACRES-B, giving eight free parameters. Thus, IHACRES-B is an extension of IHACRES-A and IHACRES-A is a special case of IHACRES-B with $p = 1$ and $l = 0$.

In the following demonstration of the framework, we begin with IHACRES-A, switching to IHACRES-B in later steps.

4.3. Application of Framework

4.3.1. DSST (Step 1)

First, we select a reference metric E and thresholds T. Given the water resources context of the case study, E should quantify the ability of the model to replicate runoff volumes, at time scales relevant to the water supply system response (days-weeks). We adopt the daily Kling-Gupta efficiency (KGE; Gupta et al., 2009) for this purpose. The KGE measures the match in mean flow, variability in flow, and timing (via linear correlation). For illustration, T_{cal} and T_{eval} are each set to 0.8, a relatively high KGE score given that KGE values may take $[-\infty, 1.0]$. Following the steps outlined in section 3:

a. *Divide* the historic record into two or more periods with contrasting climate. Given that Global Climate Model (GCM) simulations indicate likely future drying in this region (Whetton et al., 2016), model evaluation is done over the seven driest consecutive years 2005–2011 (Figure 5b), with calibration over 1970–2004.
b. *Calibrate*: IHACRES-A is calibrated using the evolutionary single-objective optimizer CMA-ES (Hansen, 2006; see supporting information, Text S4 for algorithm settings). For the objective function, we follow Fowler et al. (2018) and avoid quasi least squares metrics such as the KGE, instead adopting the Refined Index of Agreement (Willmott et al., 2012).
c. *Quantify* performance. E values are $KGE_{1970–2004}$ (E_{cal}) = 0.81; $KGE_{2005–2011}$ (E_{eval}) = 0.52.
d. *Evaluate* performance for each period by checking whether E exceeds acceptance thresholds T. In this case, $E_{cal} > T_{cal}$, but $E_{eval} < T_{eval}$. Thus, IHACRES-A fails the DSST.

4.3.2. Determine Coverage (Step 2)

The evolutionary multiobjective optimizer AMALGAM (Vrugt & Robinson, 2007) is used to test the coverage of IHACRES-A in the E_{cal}-E_{eval} space. Algorithm settings are listed in supporting information Text S4. Results are shown in Figure 6.

4.3.3. The α Acceptance Test (Step 3)

The AMALGAM output shows that IHACRES-A has no coverage of the α region. IHACRES-A thus lacks robust parameter sets—it is incapable of meeting both acceptance thresholds with the same parameter set, no matter how it is calibrated, so there is little use attempting to improve the DSST calibration method. Thus, as per Figure 3, we follow steps 4–7.

4.3.4. The β Acceptance Tests (Step 4)

The AMALGAM output shows IHACRES-A has coverage of both the β_1 region (as also confirmed in step 1) and the β_2 region. IHACRES-A is thus flexible (it can fit data it is calibrated to) but not transferable (good parameter sets in one period are poor in the other) and falls into category (b) from Figure 4.

4.3.5. Multicatchment Cross Checks (Step 5)

This step involves cross checks with nearby similar catchments. Among the Hydrologic Reference Stations, only two are within 100 km of the study catchment: 613146 (~20 km distant; area 17 km^2, mean annual

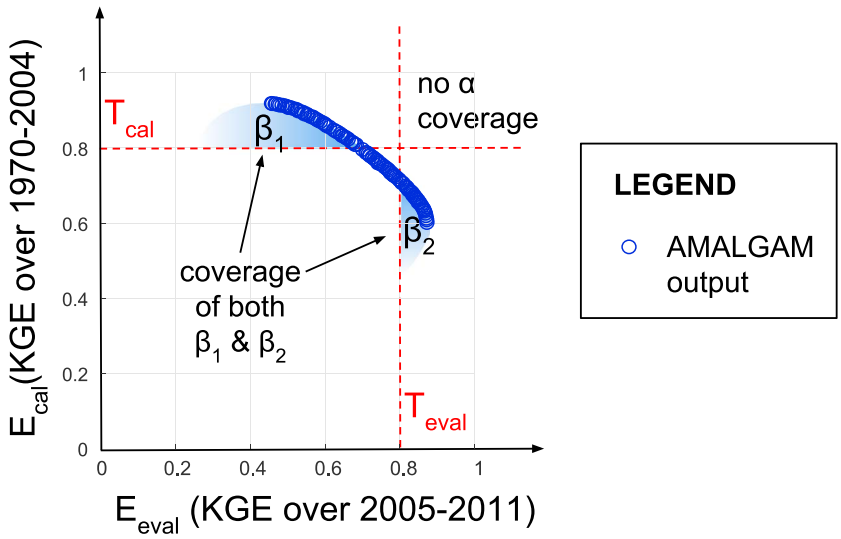


Figure 6. Coverage results (step 2) for IHACRES-A for Harvey River at Dingo Road (613002). Dark blue circles indicate AMALGAM output; light blue shading indicates implied coverage. KGE = Kling-Gupta efficiency.

Table 2
Results of Data Checks for Case Study Application, With References to Relevant Plots and Discussion in the Supporting Information

Test ID	Data check/test	Results and comments
i	Does a double mass curve (with nearby independent precipitation data) have a change in gradient?	*Test result: No* In the double mass curves in Figure S3, rainfall time series derived for this catchment is compared with rainfall derived for the four closest Hydrologic Reference Station catchments. Each curve maintains an approximately constant gradient, despite known changes in density of gauge network around 2000.
ii	Can missed events (or added events) be detected via time series/single mass curve comparisons between precipitation and runoff?	*Test result: No* Single mass curves for each year, along with time series comparisons, are shown in Figure S4. With one minor exception, all major rain events are accompanied by commensurate runoff, accounting for seasonality. No major runoff events occur without associated rainfall.
iii	Part 1: Review available information on the precipitation or runoff data. If errors seem likely conduct Part 2. Part 2: Does adding a calibratable, temporally constant scaling factor to precipitation allow coverage of the α region? Or at least, significantly improve performance?	*Test result: Significant errors considered to be relatively unlikely (Part 2 not required)* Part 1: Analysis of flow ratings and uncertainty (Text S6) reveals that 613002 has relatively high-quality flow data relative to other Hydrologic Reference Stations. Analysis of rainfall data (Text S6) indicates a relatively dense rain gauge network and relatively low spatial rainfall gradients.
iv	Does a double mass curve (with nearby independent runoff data) have a change in gradient?	*Test result: No* The runoff time series is compared among the four closest Hydrologic Reference Stations (Figure S5). Three of the four curves maintain an approximately constant gradient. The curve for station 610008 has an inflection, but further investigation reveals this is likely due to errors for 610008, not the study catchment.
v	Part 1: is there a long-term trend in a variable that is not included in the PET formulation (e.g., wind)? If yes, revise formulation and conduct part 2. Part 2: Does trialing the revised PET formulation allow coverage of the α region? Or at least, significantly improve performance?	*Test result: PET deemed unlikely to be salient cause of error* Part 1: Yes, there is a local long-term increasing trend in wind (cf. Donohue et al., 2009, p27; note that this contrasts with the general results reported by McVicar et al., 2012) and thus possible trends in evaporative demand not characterized in the adopted *Wet Environment* method (Morton, 1983). However, IHACRES-A was also deficient in other catchments with the opposite wind trend (text S11), so we did not proceed to Part 2.

Note. PET = potential evapotranspiration.

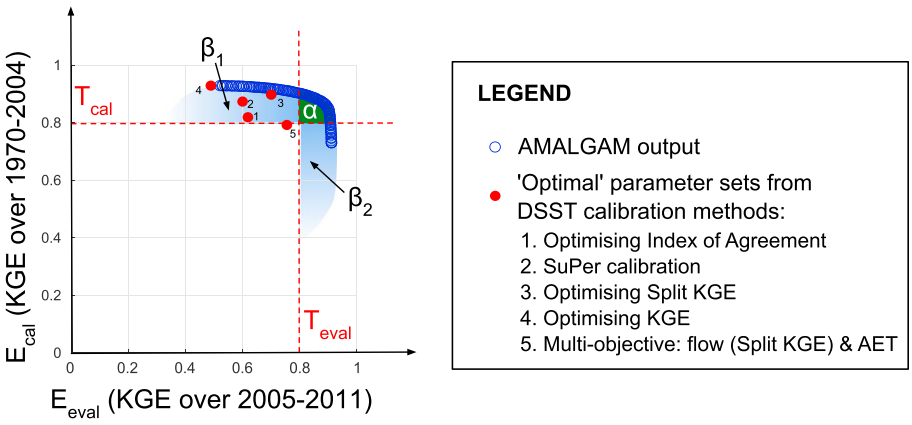


Figure 7. Coverage results (step 2) for IHACRES-B for Harvey River at Dingo Road (613002), along with parameter sets identified by different DSST calibration methods (steps 10–11). Dark blue circles indicate AMALGAM output; light blue shading indicates implied coverage. DSST = Differential Split Sample Test; KGE = Kling-Gupta efficiency; AET = Actual Evapotranspiration.

rainfall 1030 mm/year, runoff ratio 0.25) and 614044 (~50 km distant; area 73 km^2, mean annual rainfall 930 mm/year, runoff ratio 0.03). Both have broadly similar geology and land use to the study catchment, but 614044 has 8% less rainfall and 90% less runoff so is less similar hydrologically. AMALGAM results (supporting information Text S5) for 613146 match those from the study catchment relatively closely, with the same category (b) in Figure 4. Results for 614044 indicate considerable differences, with IHACRES-A having coverage of neither the β_1 region nor β_2 region, but a broadly similar shaped curve. These results are discussed further in step 7 (section 4.3.7).

4.3.6. Data Checks (Step 6)

Each data check from Table 1 is applied, with results given in Table 2 (with plots provided in the supporting information). All test results are negative, with one exception regarding the PET formulation (Morton's Wet Environment). This PET formulation does not consider wind, and a positive temporal trend in wind was reported over 1981–2006 by Donohue et al. (2009; see also McVicar et al., 2008). This is discussed further in the following step, and in supporting information Text S6.

4.3.7. Diagnostic Synthesis (Step 7)

This step considers the results of all tests in steps 4–6 and prioritizes which cause of poor performance should be remediated, be it data errors or model structural inadequacy. In this case, model structural inadequacy seems the likely salient cause. This conclusion is arrived at via multiple lines of evidence, starting with the coverage test results (Figure 6; cf. Figure 4). The replication of similar coverage results in a nearby catchment (step 5) means that one-off data errors (e.g., a deficient gauge) are not a probable cause for the poor simulations, and this is further confirmed by the uniformly negative results in tests i–iv in step 6 (Table 2). It is possible that the poor performance could be partly due to the PET formulation—note that IHACRES-A underestimates flow in the evaluation period, which is consistent with this hypothesis, as discussed in subsequent steps and supporting information Text S6. However, in this case we have access to IHACRES-A results from a wide sample of catchments (supporting information Text S11), some of which have neutral or deceasing wind trend, yet have similar coverage results. Thus, it is reasonable to prioritize model structural improvement, undertaking step 8, not step 9.

4.3.8. Structural Improvement (Step 8)

Having diagnosed model structural inadequacy, a method is now required to improve the structure. Multiple suitable methods are available in the literature (e.g., Gupta et al., 2008; Wagener et al., 2003; Westra et al., 2014). However, the focus here is on diagnosis (i.e., red steps, not green, from Figure 3), so we simply note that a multifaceted analysis of the model residuals (presented in supporting information Text S7) reveals that the two extra parameters proposed by Ye et al. (1997)—the first related to rainfall-runoff nonlinearity, and the second a threshold of runoff production—are supported by the simulation error characteristics for this case study. Thus, we adopt IHACRES-B for subsequent steps.

4.3.9. DSST (Step 1, Repeat)

This step is a repeat of section 4.3.1 using identical methods except that the model structure has changed from IHACRES-A to IHACRES-B. Updated E values are $KGE_{1970–2004}$ (E_{cal}) = 0.82 and $KGE_{2005–2011}$ (E_{eval}) = 0.62. Thus, $E_{cal} > T_{cal}$, but $E_{eval} < T_{eval}$, and IHACRES-B fails the DSST.

4.3.10. Determine Coverage (Step 2, Repeat)

Using identical methods to section 4.3.2, the coverage of IHACRES-B in the E_{cal}-E_{eval} space is defined. Results are shown in Figure 7, along with DSST results relevant to subsequent steps.

4.3.11. The α Acceptance Test (Step 3, Repeat)

The AMALGAM output shows IHACRES-B has coverage of the α region. Thus, despite the poor DSST result, no changes to the IHACRES-B model structure are needed to meet both acceptance thresholds.

4.3.12. Revise Calibration Method (Step 10)

Having confirmed coverage of the α region, the aim is now to improve the DSST calibration method so that a parameter set in this region can be identified by a repeat DSST. The new calibration method must be able to identify an α-region parameter set using *only* data from the calibration period, ie. without reference to 2005–2011. Fowler et al. (2018) recommended two objective functions for use in drying climates; having already adopted the first one (Index of Agreement), we now test the second one (Split KGE). In addition, the subperiod (SuPer) calibration method of Gharari et al. (2013) is tested (details in supporting information Text S8). Both methods work by examining multiple subperiods within the calibration period and seeking a balance between model performance across the subperiods. Calibration results are shown in Figure 7. It is seen that still $E_{val} < T_{val}$ for both methods. Results using the KGE as the objective function—a common calibration method—are also shown for reference. With so many calibration methods unable to identify a parameter set in the alpha region, it seems increasingly likely that there is insufficient information in the 1970–2004 streamflow data to correctly parameterize IHACRES-B. Thus, it is reasonable to trial an additional data type in calibration.

4.3.13. Additional Data Types (Step 11)

Finally, an extra data type is added to the DSST calibration method—namely, remotely sensed AET. Of the different data types that could be chosen here, AET is selected because a review of model fluxes indicates significant differences in the seasonal pattern of AET depending on which objective function is used, as shown in Figure 8. Calibration to KGE (parameter set 4) results in AET being highest in September, versus November for calibration to split KGE (parameter set 3). In terms of physical processes, Hughes et al.'s (2012) findings of local groundwater decline suggest sustained summer AET is possible because of plant access to groundwater, a situation more consistent with parameter set 3. Aiming for a model of maximum plausibility and hopefully increased robustness, we recalibrate IHACRES-B using a metaobjective function composed of two equally weighted components, one of which relates to AET. Specifically, the first component is the match in streamflow quantified by the Split KGE (the best performing objective function from the previous step), and the second component is the match with Moderate Resolution Imaging Spectroradiometer (MODIS)-based AET estimates of Guerschman et al. (2009), where the closeness is quantified in terms of the relative seasonal AET pattern (i.e., ignoring overall magnitude; see supporting information Text S9). The new DSST results are favorable, with E_{eval} increasing from 0.70 to 0.76 (Figure 7). Model realism increases in two ways (Figure 8): (1) a closer match with AET seasonal patterns; and (2) more plausible time series of catchment wetness, exhibiting a downward trend during the historic record that is qualitatively consistent with reported decline in groundwater (Hughes et al., 2012; Kinal & Stoneman, 2012). Trends are less realistic for the flow-only calibrations (Figure 8, parameter sets 3 and 5) because the time series oscillate close to 0 during the calibration period, giving little space for decline during the subsequent drier (evaluation) period (or a future, drier GCM scenario).

This final calibration method almost, but not quite, meets the acceptance thresholds. Thus, in this case study, we are left with the knowledge that the model structure can meet the acceptance thresholds, but we do not yet have a calibration method that can identify the acceptable parameter sets using 1970–2004 data alone.

5. Discussion

5.1. Interpretation of DSST and Coverage Tests

As already emphasized, this study makes a clear distinction between different types of model failure, and we recommend that this distinction be carried on in future practice. DSST failure (the failure of a parameterized

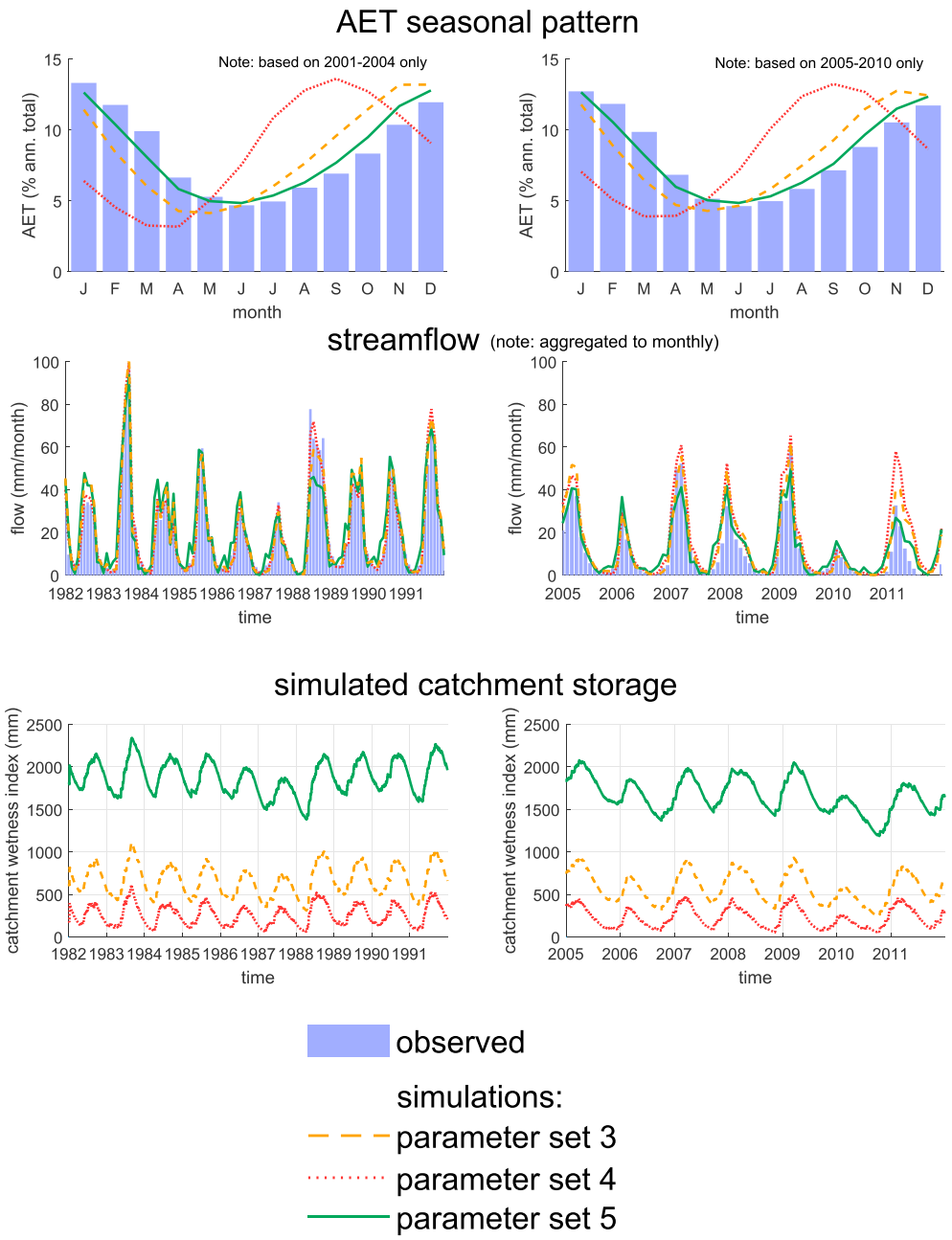
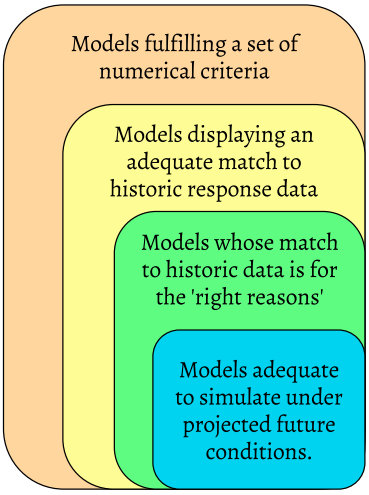


Figure 8. Observed and simulated variables for three IHACRES-B parameter sets, for calibration (left column) and evaluation (right column) periods. For an explanation of parameter sets, refer to Figure 7. Only a portion of the calibration period is shown. AET = actual evapotranspiration.

model to fulfill acceptance thresholds) was contrasted with coverage failure (the failure of every parameter set in a model structure to fulfill acceptance threshold(s)). DSST failure is contingent on the chosen calibration method, whereas coverage failure is independent of calibration method, so that failure in the latter must be due to model structural failure or data errors. In the literature, separate tests to distinguish these failure types (such as the coverage tests) are uncommon, and DSST failures are often interpreted as implying model structural failure. This can be profoundly misleading. For example, consider possible interpretations from the following DSST results from sections 4.3.1 and 4.3.9:

- 1970–2004 (calibration period) KGE: IHACRES-A: 0.81; IHACRES-B: 0.82;
- 2005–2011 (evaluation period) KGE: IHACRES-A: 0.52; IHACRES-B: 0.62.



Possible cause of inadequacy	*Practical actions for improvement*
Inadequacy of performance metrics used as criteria. Not all models with high numerical scores provide an adequate match to response data, due to limitations of performance metrics.	Refer to a wider set of criteria. Adapt criteria to the context where possible; use generic criteria sparingly. Use a variety of visual and other checks.
Right results do not imply right reasons. Not all models that provide an adequate match to response data do so for the right reasons.	Consider different response types or 'soft data'. Increase familiarity with study area(s). Dialog with experimentalists and/or locals.
The future may be unlike the past. It may deviate from past conditions sufficiently to activate unseen processes or feedbacks, cause systems to enter new states, and/or cause unexpected responses (eg. from vegetation).	Choose DSST periods that match the anticipated change. 'Space for time' methods. 'Proxy basin' split sample tests.

Figure 9. Reasons why models passing a DSST may still be inadequate to simulate runoff under projected future conditions, expressed in a hierarchical scheme. DSST = Differential Split Sample Test.

Interpreted under the false notion that DSST failure implies model structural failure, two conclusions would be as follows: (i) Both structures are unable to meet acceptance thresholds and (ii) the structural changes from Ye et al. (1997) made minimal difference to performance. As demonstrated, both conclusions are incorrect. For this reason, we strongly recommend that future studies conduct targeted tests to distinguish between failure types, such as the coverage tests suggested in this paper.

5.2. On Independence in Split Sample Testing

This framework touches on questions regarding split sample testing and what constitutes acceptable use of evaluation (often called *validation*) data. A dilemma arises because steps 10–11 (revising calibration method) involve entering a feedback loop that chooses the calibration method based (at least partially) on model performance in the evaluation period, thus compromising strict adherence to the idea of independence in split sample testing. On the other hand, not undertaking these steps when it is known that the model structure has α coverage means knowing that the model structure is capable but not allowing oneself to use this capability. As hydrologists continue in the quest for more robust simulations, this dilemma may arise more frequently and should be further debated. Furthermore, this issue should be a strong motivating factor for a well-informed initial choice of calibration method (Fowler et al., 2018; Krause & Boyle, 2005; Seibert, 2000; Yapo et al., 1996).

5.3. Are Models That Pass the DSST *Adequate*?

A common goal of DSSTs is to assess the adequacy of a model for simulation in the future, possibly under change. Is a model that passes the DSST adequate for simulation under projected future climatic conditions? Below we present multiple levels where success of a model at one level is necessary but not sufficient for success at the next level (Figure 9). Each level has distinct causes of inadequacy, as discussed below. Note that issues of uncertainty and quality in forcing data (e.g., precipitation) for future scenarios are not discussed here but are significant (e.g., Cloke et al., 2013).

Consider the set of all models (orange in Figure 9) that pass the DSST by fulfilling acceptance thresholds. No acceptance metrics can perfectly reflect the model purpose, so simulations from models in this set may be deficient over the calibration period in ways not captured by the criteria (Krause & Boyle, 2005). Various strategies may help to detect such deficiencies, including referring to a wider set of criteria (Gupta et al., 1998, 2008; Thirel et al., 2015) and using a variety of checks and visualizations (Bennett et al., 2013; Thirel et al., 2015). It is important to select criteria that are closely matched to the model purpose, as models that are adequate for one type of application may not be adequate for other types, since they are often unable to match different aspects of the flow regime simultaneously (e.g., high flows and low flows with the same parameter set).

Moving to the next level in Figure 9, models that display an adequate match (however this is defined) in numerical outputs may not do so for the right reasons (Beven, 2006; Kirchner, 2006; Klemeš, 1986). For example, a model may activate a process that experimental data or site experience reveals is the wrong mechanism for the catchment (e.g., Hortonian flow in a catchment with high infiltration). Strategies for revealing this kind of inadequacy include site familiarization/seeking local knowledge (Holländer et al., 2014), more difficult tests through extra data (e.g., Mroczkowski et al., 1997), and dialogue between modelers and experimental hydrologists (Freer et al., 2004; Seibert & McDonnell, 2002). All of these may help to discriminate between parameter sets that are equifinal with respect to numerical output. We note that the meaning of *right reasons* may vary depending on the underlying perceptual model and philosophical viewpoint of the modeler (e.g., Gupta et al., 2012).

Finally, even models that match historic data for the right reasons (green) may not be adequate to simulate under projected change (blue) because the future may be so different from the past as to change the mechanisms that govern the rainfall-runoff relationship (cf. Saft et al., 2015). New processes may become dominant, and living components of the system (e.g., vegetation) may respond unexpectedly due to complex feedbacks (Curtis & Wang, 1998; D'Odorico & Porporato, 2004; Rodriguez-Iturbe et al., 1999) possibly leading to less resilience following disturbances (Peterson et al., 2009). The concept known as *trading space for time* (Peel & Blöschl, 2011) may be useful, whether for model parameterization (Singh et al., 2011) or to test parameterized models in climatic conditions beyond the study area's historic record by using another catchment entirely (e.g., proxy basin split sample tests; Klemeš, 1986; Refsgaard et al., 2014). One risk with such methods is that, assuming catchments coevolve, they reflect the actions of processes over centuries or millennia, in contrast to climate change which is likely to occur relatively quickly. An additional confounding factor is that records of catchment hydroclimatology are themselves subject to increasing CO_2 concentrations (Roderick et al., 2015). In general, our ability to successfully transition from the green category to the blue is limited because our knowledge of the future is fundamentally incomplete. Thus, although the DSST may be *the best possible evaluation method* (Refsgaard et al., 2014), the adequacy of models that pass the DSST is far from guaranteed.

5.4. Further Research Regarding Calibration Methods

The case study demonstrates the importance of calibration methods in modeling outcomes, and further research is recommended in this area. As stated, choice of calibration method (including selection of objective function) should be grounded in hydrological theory and aim to select parameters with high fidelity to dominant processes. Achieving this will likely require consideration of typical errors present in training and forcing data (e.g., Coxon et al., 2014; McMillan et al., 2012; Schoups & Vrugt, 2010; Sorooshian & Dracup, 1980), and such considerations were shown by Fowler et al. (2018) to yield significantly improved DSST results. Practically, one potential problem for the modeler is that there are many studies suggesting improved calibration and sampling methods but little guidance to choose between them. For example, a water resource-focused climate change impact study could potentially use many methods, each with strong supporting theory which have been developed or demonstrated for changing climates. Methods include (i) those that use calibration data differently by defining wetter and drier subperiods of the calibration period (Choi & Beven, 2007; Freer et al., 2003; Gharari et al., 2013); (ii) trading space for time approaches which incorporate information from other catchments in the same region to predict hydrologic response (e.g., Singh et al., 2011); (iii) alternative objective functions that place different focus on hydrological behaviors, for example, applying transforms on flow values before sum of squares calculations or using sum of absolute errors rather than sum of squares (Fowler et al., 2018, see also Krause & Boyle, 2005; Willmott et al., 2012). Furthermore, methods that identify ensembles that are likely to be more robust to change such as (iv) data depth approaches (Bárdossy & Singh, 2008) and (v) limits of acceptability approaches (Liu et al., 2009) may also be relevant to the context of changes in climatic conditions. In response to this prolific variety of methods, we recommend three classes of study to increase the value of existing work: (a) intermethod comparison, to guide method choice; (b) studies combining ideas from different methods—for example, methods (iii) and (iv) could be gainfully combined; and (c) studies testing calibration methods in regions with relatively high hydroclimatic variability (e.g., the *crash tests* of Coron et al., 2012), to provide assurance that the methods work successfully on the most challenging available data, while acknowledging that this does not guarantee adequacy for future conditions, as discussed above.

5.5. Generality of the Framework

It is noted that the framework can be easily adapted for multiple catchments, to serve the needs of large-sample hydrology studies (Gupta et al., 2014). For example, supporting information Text S10/Figure S9 give an example across multiple catchments from Australia, showing how analysis can be meaningfully summarized across large samples of catchments. Further, we argue that most elements of the framework are applicable beyond simple conceptual models and climate change studies. For example, the principle of the *coverage check* as a first step following split sample failure (followed by updating the calibration method in the case of alpha coverage) is generally relevant to all applications of split sample testing, including changes across space rather than time, for changing land use rather than changing climate, and/or to process-based as well as conceptual models. We are currently applying the framework to stress test a gridded continental-scale hydrological model, using consecutive coverage tests while sequentially adding catchments to discover at which point the model equations can no longer perform across the required spatial domain, and we look forward to reporting these results in a future article.

6. Conclusions and Recommendations

This paper presents a framework to improve rainfall-runoff simulations under a changing climate. Model evaluation based solely on the DSST is hampered due to contingency on the chosen calibration method, and it is difficult to distinguish which cases of DSST failure are truly caused by model structural inadequacy. The proposed framework addresses this problem by diagnosing the salient cause of poor model performance, be it structural inadequacy, poor calibration method, or data errors. We demonstrated methods to explore the limits of model robustness and flexibility over multiple climatic periods, which can be used to discriminate between these causes and prioritize remedial action.

Using a case study catchment from Australia, improved inference of structural failure and clearer evaluation of model structural improvements were demonstrated. Interpreted under the false notion that DSST failure implies model structural failure, the modeler would have wrongly concluded from our results that the model structural improvements made minimal difference to model performance in the case study. In contrast, the coverage tests in the framework demonstrated an enhanced robustness previously hidden because the wrong parameter set was initially chosen in the DSST.

In the literature, DSST failure is often assumed to imply model structural failure, and the above result shows that this assumption can be profoundly misleading, possibly leading to erroneous comparisons between candidate model structures and inability to properly assess the benefit of structural improvements. It is recommended that future studies use this framework to (i) identify cases where poor simulations are due to poor parameterization or data errors, remediating these cases without recourse to structural changes and (ii) use the remaining cases to gain greater clarity into what structural changes are needed to improve model performance in changing climate.

Acknowledgments
The authors gratefully acknowledge the support of the Australian Government in carrying out this work. Specifically, Keirnan Fowler's work was supported by an Australian Postgraduate Award, and Murray Peel is the recipient of an Australian Research Council Future Fellowship (FT120100130). Authors from the University of Bristol acknowledge the support of the UK's Natural Environment Research Council grant MaRIUS: Managing the Risks, Impacts and Uncertainties of droughts and water Scarcity (NE/L010399/1). Streamflow data used in this project were from the Australian Bureau of Meteorology's (BOM) Hydrologic Reference Station project website (Turner, 2012), www.bom.gov.au/water/hrs. Rainfall data were from the Australian Water Availability Project (AWAP) project (Jones et al., 2009), www.bom.gov.au/jsp/awap/. Potential evapotranspiration data were from the SILO project (Jeffrey et al., 2001), www.longpaddock.qld.gov.au/silo/. Information on historical bushfires was made available by the Department of Biodiversity, Conservation and Attractions, Western Australia. MODIS-based AET estimates based on Guerschman et al. (2009) were made available by the Australian Bureau of Meteorology. We thank Shervan Gharari, Luis Samaniego, and two anonymous reviewers for their constructive feedback on this work.

References

Ambroise, B., Beven, K., & Freer, J. (1996). Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity. *Water Resources Research*, *32*(7), 2135–2145. https://doi.org/10.1029/95WR03716

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M. H., Oudin, L., Mathevet, T., et al. (2012). All that glitters is not gold: The case of calibrating hydrological models. *Hydrological Processes*, *26*(14), 2206–2210. https://doi.org/10.1002/hyp.9264

Bárdossy, A., & Singh, S. K. (2008). Robust estimation of hydrological model parameters. *Hydrology and Earth System Sciences*, *12*(6), 1273–1283. https://doi.org/10.5194/hess-12-1273-2008

Beck, M. B. (2002). *Environmental foresight and models: A manifesto* (Vol. 22). Athens, Georgia: Gulf Professional Publishing.

Beck, M. B. (2005). Environmental foresight and structural change. *Environmental Modelling & Software*, *20*(6), 651–670. https://doi.org/10.1016/j.envsoft.2004.04.005

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environmental Modelling and Software*, *40*, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Bergström, S., Carlsson, B., Gardelin, M., Lindström, G., Petterson, A., & Rummukainen, M. (2001). Climate change impacts on runoff in Sweden—Assessments by global climate models, dynamical downscalling and hydrological modelling. *Climate Research*, *16*(2), 101–112. https://doi.org/10.3354/cr016101

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K., Smith, P. J., & Wood, A. (2011). On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, *15*(10), 3123–3133. https://doi.org/10.5194/hess-15-3123-2011

Beven, K., & Westerberg, I. (2011). On red herrings and real herrings: Disinformation and information in hydrological inference. *Hydrological Processes*, *25*(10), 1676–1680. https://doi.org/10.1002/hyp.7963

Brigode, P., Oudin, L., & Perrin, C. (2013). Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change? *Journal of Hydrology*, *476*, 410–425. https://doi.org/10.1016/j.jhydrol.2012.11.012

Broderick, C., Matthews, T., Wilby, R. L., Bastola, S., & Murphy, C. (2016). Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. *Water Resources Research*, *52*, 8343–8373. https://doi.org/10.1002/2016WR018850

Chiew, F., Whetton, P., McMahon, T., & Pittock, a. B. (1995). Simulation of the impacts of climate change on runoff and soil moisture in Australian catchments. *Journal of Hydrology*, *167*(1–4), 121–147. https://doi.org/10.1016/0022-1694(94)02649-V

Chiew, F. H. S., Teng, J., Vaze, J., Post, D. A., Perraud, J. M., Kirono, D. G. C., & Viney, N. R. (2009). Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method. *Water Resources Research*, *45*, W10414. https://doi.org/10.1029/2008WR007338

Choi, H. T., & Beven, K. (2007). Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework. *Journal of Hydrology*, *332*(3–4), 316–336. https://doi.org/10.1016/j.jhydrol.2006.07.012

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The effects of climate change on the hydrology and water resources of the Colorado River basin. *Climatic Change*, *62*(1-3), 337–363. https://doi.org/10.1023/B:CLIM.0000013684.13621.1f

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). A unified approach for process-based hydrologicmodeling: 1. Modeling concept. *Water Resources Research*, *51*, 1–17. https://doi.org/10.1002/2015WR017200.A

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., et al. (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources*, *31*(10), 1309–1324. https://doi.org/10.1016/j.advwatres.2008.06.005

Cloke, H. L., Wetterhall, F., He, Y., Freer, J. E., & Pappenberger, F. (2013). Modelling climate impact on floods with ensemble climate projections. *Quarterly Journal of the Royal Meteorological Society*, *139*(671), 282–297. https://doi.org/10.1002/qj.1998

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., & Hendrickx, F. (2014). On the lack of robustness of hydrologic models regarding water balance simulation: A diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments. *Hydrology and Earth System Sciences*, *18*(2), 727–746. https://doi.org/10.5194/hess-18-727-2014

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*, W05552. https://doi.org/10.1029/2011WR011721

Covey, C., AchutaRao, K. M., Cubasch, U., Jones, P., Lambert, S. J., Mann, M. E., et al. (2003). An overview of results from the Coupled Model Intercomparison Project. *Global and Planetary Change*, *37*(1–2), 103–133. https://doi.org/10.1016/S0921-8181(02)00193-5

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes*, *28*(25), 6135–6150. https://doi.org/10.1002/hyp.10096

Curtis, P. S., & Wang, X. (1998). A meta-analysis of elevated CO_2 effects on woody plant mass, form, and physiology. *Oecologia*, *113*(3), 299–313. https://doi.org/10.1007/s004420050381

de Vos, N. J., Rientjes, T. H. M., & Gupta, H. V. (2010). Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering. *Hydrological Processes*, *24*(20), 2840–2850. https://doi.org/10.1002/hyp.7698

D'Odorico, P., & Porporato, A. (2004). Preferential states in soil moisture and climate dynamics. *Proceedings of the National Academy of Sciences*, *101*(24), 8848–8851. https://doi.org/10.1073/pnas.0401428101

Donohue, R. J., Mcvicar, T. R., & Roderick, M. L. (2009). Generating Australian potential evaporation data suitable for assessing the dynamics in evaporative demand within a changing climate, Tech. Rep. *December*, CSIRO.

Faramarzi, M., Abbaspour, K. C., Ashraf Vaghefi, S., Farzaneh, M. R., Zehnder, A. J. B., Srinivasan, R., & Yang, H. (2013). Modeling impacts of climate change on freshwater availability in Africa. *Journal of Hydrology*, *480*, 85–101. https://doi.org/10.1016/j.jhydrol.2012.12.016

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*, W11510. https://doi.org/10.1029/2010WR010174

Forster, P., Ramaswamy, V., Artaxo, P., Berntsen, T., Betts, R., Fahey, D. W., et al. (2007). Changes in atmospheric constituents and in radiative forcing. In S. Solomon, et al. (Eds.), *Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 129–234). Cambridge, United Kingdom and New York, NY: Cambridge University Press.

Forzieri, G., Feyen, L., Rojas, R., Flörke, M., Wimmer, F., & Bianchi, A. (2014). Ensemble projections of future streamflow droughts in Europe. *Hydrology and Earth System Sciences*, *18*(1), 85–108. https://doi.org/10.5194/hess-18-85-2014

Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved rainfall-runoff calibration for drying climate: Choice of objective function. *Water Resources Research*, *54*, 3392–3408. https://doi.org/10.1029/2017WR022466

Fowler, K. J. A. (2017). Towards improved rainfall runoff modelling in changing climatic conditions, PhD thesis, University of Melbourne, Department of Infrastructure Engineering.

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., & Peterson, T. J. (2016). Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. *Water Resources Research*, *52*, 1820–1846. https://doi.org/10.1002/2015WR018068

Freer, J., Beven, K., & Peters, N. (2003). Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure. In Q. Duan, H. Gupta, S. Sorooshian, A. Rousseau, & R. Turcotte (Eds.), *Calibration of watershed models* (pp. 69–87). Washington, DC: American Geophysical Union. https://doi.org/10.1029/WS006p0069

Freer, J., McMillan, H., McDonnell, J., & Beven, K. (2004). Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology*, *291*(3), 254–277. https://doi.org/10.1016/j.jhydrol.2003.12.037, catchment modelling: Towards an improved representation of the hydrological processes in real-world model applications.

Gallant, J., Dowling, T., Read, A. M., Wilson, N., Tickle, P., & Inskeep, C. (2011). 1 second SRTM derived products user guide, *Tech. Rep. October*, Geoscience Australia.

Gharari, S., Hrachowitz, M., Fenicia, F., & Savenije, H. H. G. (2013). An approach to identify time consistent model parameters: Sub-period calibration. *Hydrology and Earth System Sciences*, *17*(1), 149–161. https://doi.org/10.5194/hess-17-149-2013

Guerschman, J. P., Dijk, A. I. V., Mattersdorf, G., Beringer, J., Hutley, L. B., Leuning, R., et al. (2009). Scaling of potential evapotranspiration with MODIS data reproduces flux observations and catchment water balance observations across Australia. *Journal of Hydrology*, *369*(1-2), 107–119. https://doi.org/10.1016/j.jhydrol.2009.02.013

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*, W08301. https://doi.org/10.1029/2011WR011044

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, *34*(4), 751–763. https://doi.org/10.1029/97WR03495

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Hadka, D., & Reed, P. M. (2013). Borg: An auto-adaptive many-objective evolutionary computing framework. *Evolutionary Computation*, *21*(2), 231–259. https://doi.org/10.1162/EVCO_a_00075

Hansen, N. (2006). The CMA evolution strategy: A comparing review. *Studies in Fuzziness and Soft Computing*, *192*(2006), 75–102. https://doi.org/10.1007/11007937-4

Holländer, H., Bormann, H., Blume, T., Buytaert, W., Chirico, G., Exbrayat, J.-F., et al. (2014). Impact of modellers' decisions on hydrological a priori predictions. *Hydrology and Earth System Sciences*, *18*(6), 2065–2085. https://doi.org/10.5194/hess-18-2065-2014

Hughes, J. D., Petrone, K. C., & Silberstein, R. P. (2012). Drought, groundwater storage and stream flow decline in southwestern Australia. *Geophysical Research Letters*, *39*, L03408. https://doi.org/10.1029/2011GL050797

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, *29*(8), 2637–2649. https://doi.org/10.1029/93WR00877

Jakeman, A. J., Littlewood, I. G., & Whitehead, P. G. (1990). Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology*, *117*(1–4), 275–300. https://doi.org/10.1016/0022-1694(90)90097-H

Jeffrey, S. J., Carter, J. O., Moodie, K. B., & Beswick, A. R. (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling and Software*, *16*(4), 309–330. https://doi.org/10.1016/S1364-8152(01)00008-1

Jones, D. A., Wang, W., & Fawcett, R. (2009). High-quality spatial climate data-sets for Australia. *Australian Meteorological and Oceanographic Journal*, *58*(04), 233–248. https://doi.org/10.22499/2.5804.003

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, *42*, W03407. https://doi.org/10.1029/2005WR004368

Kelleher, C. A., & Shaw, S. B. (2018). Is ET often oversimplified in hydrologic models? Using long records to elucidate unaccounted for controls on ET. *Journal of Hydrology*, *557*, 160–172. https://doi.org/10.1016/j.jhydrol.2017.12.018

Kinal, J., & Stoneman, G. (2012). Disconnection of groundwater from surface water causes a fundamental change in hydrology in a forested catchment in south-western Australia. *Journal of Hydrology*, *472*, 14–24.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Krause, P., & Boyle, D. P. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *5*, 89–89), 97. https://doi.org/10.5194/adgeo-5-89-2005

Krysanova, V., Vetter, T., Eisner, S., Huang, S., Pechlivanidis, I., Strauch, M., et al. (2017). Intercomparison of regional-scale hydrological models and climate change impacts projected for 12 large river basins worldwide—A synthesis. *Environmental Research Letters*, *12*(10), 105,002. https://doi.org/10.1088/1748-9326/aa8359

Liu, Y., Freer, J., Beven, K., & Matgen, P. (2009). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, *367*(1-2), 93–103. https://doi.org/10.1016/j.jhydrol.2009.01.016

Ljung, L. (1998). System identification. In *Signal analysis and prediction* (pp. 163–173). New York: Springer.

Marshall, L., Nott, D., & Sharma, A. (2007). Towards dynamic catchment modelling: A Bayesian hierarchical mixtures of experts framework. *Hydrological Processes*, *21*(7), 847–861. https://doi.org/10.1002/hyp.6294

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384

McVicar, T. R., Roderick, M. L., Donohue, R. J., Li, L. T., Van Niel, T. G., Thomas, A., et al. (2012). Global review and synthesis of trends in observed terrestrial near-surface wind speeds: Implications for evaporation. *Journal of Hydrology*, *416-417*, 182–205. https://doi.org/10.1016/j.jhydrol.2011.10.024

McVicar, T. R., Van Niel, T. G., Li, L. T., Roderick, M. L., Rayner, D. P., Ricciardulli, L., & Donohue, R. J. (2008). Wind speed climatology and trends for Australia, 1975–2006: Capturing the stilling phenomenon and comparison with near-surface reanalysis output. *Geophysical Research Letters*, *35*, L20403. https://doi.org/10.1029/2008GL035627

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, H. M., et al. (2007). Global climate projections. In S. Solomon, et al. (Eds.), *Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 747–846). Cambridge, United Kingdom and New York, NY: Cambridge University Press.

Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research*, *47*, W02531. https://doi.org/10.1029/2010WR009505

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Zbigniew, W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity is dead: Whither water management? *Science*, *319*(5863), 573–574. https://doi.org/10.1126/science.1151915

Montanari, A., Young, G., Savenije, H., Hughes, D., Wagener, T., Ren, L., et al. (2013). Panta Rhei—"Everything flows: Change in hydrology and society" The IAHS Scientific Decade 2013–2022. *Hydrological Sciences Journal*, *58*(6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088

Morton, F. I. (1983). Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology. *Journal of Hydrology*, *66*(1-4), 1–76. https://doi.org/10.1016/0022-1694(83)90177-4

Mroczkowski, M., Raper, G. P., & Kuczera, G. (1997). The quest for more powerful validation of conceptual catchment models. *Water Resources Research*, *33*(10), 2325–2335. https://doi.org/10.1029/97WR01922

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

Pechlivanidis, I. G., Arheimer, B., Donnelly, C., Hundecha, Y., Huang, S., Aich, V., et al. (2017). Analysis of hydrological extremes at different hydro-climatic regimes under present and future conditions. *Climatic Change*, *141*(3), 467–481. https://doi.org/10.1007/s10584-016-1723-0

Peel, M. C., & Blöschl, G. (2011). Hydrological modelling in a changing world. *Progress in Physical Geography*, *35*(2), 249–261. https://doi.org/10.1177/0309133311402550

Peterson, T. J., Argent, R. M., Western, A. W., & Chiew, F. H. S. (2009). Multiple stable states in hydrological models: An ecohydrological investigation. *Water Resources Research*, *45*, W03406. https://doi.org/10.1029/2008WR006886

Petrone, K. C., Hughes, J. D., Van Niel, T. G., & Silberstein, R. P. (2010). Streamflow decline in southwestern Australia, 1950–2008. *Geophysical Research Letters*, *37*, L11401. https://doi.org/10.1029/2010GL043102

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, *32*(7), 2189–2202. https://doi.org/10.1029/96WR00896

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., et al. (2014). A framework for testing the ability of models to project climate change and its impacts. *Climatic Change*, *122*(1–2), 271–282. https://doi.org/10.1007/s10584-013-0990-2

Reichert, P., & Omlin, M. (1997). On the usefulness of overparameterized ecological models. *Ecological Modelling*, *95*(2–3), 289–299. https://doi.org/10.1016/S0304-3800(96)00043-9

Roderick, M. L., Greve, P., & Farquhar, G. D. (2015). On the assessment of aridity with changes in atmospheric CO_2. *Water Resources Research*, *51*, 5450–5463. https://doi.org/10.1002/2015WR017031

Rodriguez-Iturbe, I., D'Odorico, P., Porporato, A., & Ridolfi, L. (1999). On the spatial and temporal links between vegetation, climate, and soil moisture. *Water Resources Research*, *35*(12), 3709–3722. https://doi.org/10.1029/1999WR900255

Saft, M., Peel, M. C., Western, A. W., Perraud, J.-M., & Zhang, L. (2016). Bias in streamflow projections due to climate-induced shifts in catchment response. *Geophysical Research Letters*, *43*, 1574–1581. https://doi.org/10.1002/2015GL067326

Saft, M., Western, A. W., Zhang, L., Peel, M. C., & Potter, N. J. (2015). The influence of multiyear drought on the annual rainfall-runoff relationship: An Australian perspective. *Water Resources Research*, *51*, 2444–2463. https://doi.org/10.1002/2014WR015348

Samaniego, L., Kumar, R., Breuer, L., Chamorro, A., Flörke, M., Pechlivanidis, I. G., et al. (2017). Propagation of forcing and model uncertainties on to hydrological drought characteristics in a multi-model century-long experiment in large river basins. *Climatic Change*, *141*(3), 435–449. https://doi.org/10.1007/s10584-016-1778-y

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*, W10531. https://doi.org/10.1029/2009WR008933

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000

Seibert, J. (2003). Reliability of model predictions outside calibration conditions. *Nordic Hydrology*, *34*(5), 477–492. https://doi.org/10.2166/nh.2003.028

Seibert, J., & McDonnell, J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multi-criteria model calibration. *Water Resources Research*, *38*(11), 1241. https://doi.org/10.1029/2001WR000978

Seiller, G., & Anctil, F. (2015). How do potential evapotranspiration formulas influence hydrological projections? *Hydrological Sciences Journal*, *61*(12), 2249–2266. https://doi.org/10.1080/02626667.2015.1100302

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Singh, R., van Werkhoven, K., & Wagener, T. (2014). Hydrological impacts of climate change in gauged and ungauged watersheds of the Olifants basin: A trading-space-for-time approach. *Hydrological Sciences Journal*, *59*(1), 29–55. https://doi.org/10.1080/02626667.2013.819431

Singh, R., Wagener, T., Van Werkhoven, K., Mann, M. E., & Crane, R. (2011). A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate-accounting for changing watershed behavior. *Hydrology and Earth System Sciences*, *15*(11), 3591–3603. https://doi.org/10.5194/hess-15-3591-2011

Smith, A., Bates, P., Freer, J., & Wetterhall, F. (2014). Investigating the application of climate models in flood projection across the Uk. *Hydrological Processes*, *28*(5), 2810–2823. https://doi.org/10.1002/hyp.9815

Smith, A., Freer, J., Bates, P., & Sampson, C. (2014). Comparing ensemble projections of flooding against flood estimation by continuous simulation. *Journal of Hydrology*, *511*, 205–219. https://doi.org/10.1016/j.jhydrol.2014.01.045

Sorooshian, S., & Dracup, J. A. (1980). Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases. *Water Resources Research*, *16*(2), 430–442. https://doi.org/10.1029/WR016i002p00430

Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P., et al. (2015). Hydrology under change: An evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrological Sciences Journal - Journal des Sciences Hydrologiques*, *60*, 7–8. https://doi.org/10.1080/02626667.2014.967248

Turner, M. (2012). Hydrologic Reference Station selection guidelines.

van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., De Jeu, R. A. M., Liu, Y. Y., Podger, G. M., et al. (2013). The millennium drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, *49*, 1040–1057. https://doi.org/10.1002/wrcr.20123

van Werkhoven, K., Wagener, T., Reed, P., & Tang, Y. (2008). Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resources Research*, *44*, W01429. https://doi.org/10.1029/2007WR006271

Vaze, J., Chiew, F. H. S., Perraud, J. M., Viney, N., Post, D., Teng, J., et al. (2010). Rainfall-runoff modelling across southeast Australia: Datasets, models and results. *Australian Journal of Water Resources*, *14*(2), 101–116.

Vogel, R. M., Wilson, I., & Daly, C. (1999). Regional regression models of annual streamflow for the United States. *Journal of Irrigation and Drainage Engineering*, *125*(3), 148–157. https://doi.org/10.1061/(ASCE)0733-9437(1999)125:3(148)

Vrugt, J. a., & Robinson, B. a. (2007). Improved evolutionary optimization from genetically adaptive multimethod search. *Proceedings of the National Academy of Sciences*, *104*(3), 708–711. https://doi.org/10.1073/pnas.0610471104

Vrugt, J. a., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. a. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*, W00B09. https://doi.org/10.1029/2007WR006720

Wagener, T. (2003). Evaluation of catchment models. *Hydrological Processes*, *17*(16), 3375–3378. https://doi.org/10.1002/hyp.5158

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., & Gupta, H. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrological Processes*, *17*(2), 455–476. https://doi.org/10.1002/hyp.1135

Westra, S., Thyer, M., Leonard, M., Kavetski, D., & Lambert, M. (2014). A strategy for diagnosing and interpreting hydrological model nonstationarity. *Water Resources Research*, *50*, 1–24. https://doi.org/10.1002/2013WR014719

Whetton, P. H., Grose, M. R., & Hennessy, K. J. (2016). A short history of the future: Australian climate projections 1987–2015. *Climate Services*, *2-3*, 1–14. https://doi.org/10.1016/j.cliser.2016.06.001

Wilby, R. L. (2005). Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrological Processes*, *19*(16), 3201–3219. https://doi.org/10.1002/hyp.5819

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, *32*(13), 2088–2094. https://doi.org/10.1002/joc.2419

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, *181*(1-4), 23–48. https://doi.org/10.1016/0022-1694(95)02918-4

Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M., & Jakeman, A. J. (1997). Performance of conceptual rainfall-runoff models in low yielding ephemeral catchments. *Water Resources Research*, *33*(1), 153–166. https://doi.org/10.1029/96WR02840